DOCUMENT RESUME

ED 181 027 TM 009 784

AUTHOR Haney, Walt; Strenio, Andrew
TITLE Summary of the Spring 1979 Conference of the National Consortium on Testing, (April 30-May 1, 1979).
INSTITUTION Huron Inst., Cambridge, Mass.
PUB DATE 1 Jun 79
NOTE 83p.; Three appendices were removed before filming

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Conference Reports; Confidentiality; Educational Assessment; Information Needs; Item Banks; *Minimum Competency Testing; Occupational Tests; Parent Role; Program Evaluation; Standardized Tests; State Programs; *Student Testing; Test Construction; *Testing Problems; Testing Programs
IDENTIFIERS *National Consortium on Testing; *Test Coaching; Test Security

ABSTRACT

A summary, prepared from notes and tape recordings, is presented of the 1979 conference of the National Consortium on Testing. Panel descussions on the following topics are included: (1) introduction; (2) update--recent and upcoming events concerning testing; (3) test preparation and coaching; (4) minimum competency testing; (5) program evaluation and testing; (6) viewing and discussion of public education films on testing; (7) topical sessions, including information on testing for parents and the public, testing and program evaluation, and minimum competency testing; (8) open versus secure testing; (9) publicly available test item banks; and (10) open competence testing. Two articles, "A Study of Minimum Competency Testing Programs, and "Why Should Some Tests be Secure," are appended. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED181027

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SUMMARY OF THE SPRING 1979 CONFERENCE

OF THE

NATIONAL CONSORTIUM ON TESTING

April 30 - May 1, 1979

Prepared by

Walt Haney

and

Andrew Strenio

Huron Institute
123 Mount Auburn Street
Cambridge, Mass. 02138
617/661-9285

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Paul L. Houts

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

June 1, 1979

TM009 784

This summary has been prepared from notes and tape recordings of the Consortium conference. It does not, however, represent a verbatim account of the conference.

TABLE OF CONTENTS

	page:
I Introduction	1
Vito Perrone Walt Haney	
II Update - Recent and Upcoming Events Concerning Testing	4
Josephine Barresi Judah Schwartz John Maxwell David Seeley Frances Quinto James Ward Walt Haney (for Deborah Meier) Ann Kahn Hayes Mizell Gregory Kelley Barbara Lieb-Brilhart Susan Brown Vito Perrone Eric Gardner James Alatis	
III Test Preparation and Coaching	13
Stephen Ivens Arthur Levine Discussion	
IV Minimum Competency Testing	19
Corinne Scott William Gorth Discussion Hayes Mizell Discussion	
V Program Evaluation and Testing	25
Daniel Stufflebeam Walt Haney Discussion	
VI Viewing and Discussion of Public Education Films on Testing	33

VII Topical Sessions 35

Information on testing for Parents and the Public
Testing and Program Evaluation
Minimum Competency Testing

VIII Open Versus Secure Testing 37

William Corbett
Alice Irby
Steve Solomon
Winsor Lott
Discussion

IX Publicly Available Test Item Banks 47

Jerrold Zacharias
Discussion

X Open Competence Testing 50

David McClelland
Discussion

Attachments 55

1. "The Standardized Test Contributes to the Furor" article from PTA Today

2. TESOL Memo "Concerns About Statewide Competency Testing"

3. "A Study of Minimum Competency Testing Programs" description of an NIE-funded project conducted by National Evaluation Systems of Amherst, Massachusetts

4. Handout Concerning Draft Standards for Educational Evaluation

5. "Why Should Some Tests Be Secure" by John Fremer and Alice Irby, Educational Testing Service

I. INTRODUCTION

The spring 1979 Membership conference of the National Consortium on Testing was held April 30 - May 1, Monday and Tuesday, at the Hyatt House in Arlington, Virginia. This paper summarizes the general sessions of the conference on Monday and Tuesday. The business meeting, held Tuesday afternoon, will be reported separately.

Consortium President Vito Perrone opened the meeting on Monday morning with comments on the flooding of the Red River in North Dakota and Minnesota, and on "educational" consequences of the flood. He welcomed Consortium representatives and other guests to the spring meeting, and in particular welcomed the eight organizations which joined the Consortium since the last membership conference:

1. The American College Testing Program
2. Association of American Medical Colleges
   (Division of Educational Measurement and Research)
3. Chicano Education Project
4. The College Board
5. Division for Study and Research in Education (MIT)
6. National Association for the Education of Young Children
7. Southeastern Public Education Program (AFSC)
8. Teachers of English to Speakers of Other Languages

Vito noted that the Consortium had its origins in 1974-75 when 23 educational associations came together to share some of their concerns over the growing role of testing in society and particularly in the schools, A resolution came from one such early meeting, calling for greater public education about tests and testing, for more public

disclosure about the content of tests, and the development of helpful alternative assessment procedures. Since those early meetings, Vito continued, the Consortium has expanded at least twofold, and developed a formal organizational structure. The purposes of the Consortium are fairly straightforward, and, he mentioned, are outlined in a brief flyer describing the Consortium. In order to pursue its purposes, the Consortium has enlisted the Huron Institute to provide staff support and research, which, Vito suggested, have so far proved enormously constructive. Through meetings and information exchange, people involved with the Consortium have enlarged their understanding of testing issues and the knowledge base concerning testing.

The sessions on Monday and Tuesday, Vito suggested, hold particular promise in large measure because issues raised in the 1974-75 period are being highlighted. These issues, such as the debate over open versus secure testing, test coaching, minimum competency testing, and public education concerning testing are gaining widespread attention. The focus on these issues in the spring 1979 conference has grown out of suggestions and critical observations of Consortium members, and Vito suggested, our understanding of them will be further enhanced by critical attention and questions of people attending the conference.

Walt Haney reported briefly on the Huron Institute work for the Consortium. He reiterated Vito's welcome to the new member organizations of the Consortium and noted that several other groups have already expressed interest in joining the Consortium --

an apparent sign of organizational health for the group. Huron continues to distribute information from the Consortium including, both meeting summaries and staff circulars, and the mailing list for these materials has now reached nearly 300. Huron staff have also attended several meetings of other organizations to describe the Consortium and its concerns. On substantive matters, work at Huron is proceeding in the four areas outlined in Huron's original program plans. In the area of Truth-in-Testing, Huron staff are completing a review of most commonly used standardized tests and are undertaking a small pilot study of a random sample of test items in order to illuminate issues of test item quality. With respect to the federal government's role in testing, Walt pointed attention to the sessions at the spring conference concerning minimum competency testing and USOE's proposed models for evaluating Title I programs. Concerning state and local testing policies, Walt noted that the two staff circulars Huron has prepared concerning minimum competency testing have proven particularly popular. With regard to alternative assessment procedures, he remarked that this was a major topic of attention at the fall 1978 Consortium conference and that two presentations at the current conference -- by Jerrold Zacharias and by David McClelland -- would be dealing with ideas for alternative means of assessment. Also the issue of alternatives will be receiving additional attention at future meetings

and in Huron work for the Consortium. In closing, Walt mentioned that with the sanction of the Consortium's Board of Directors, Huron is beginning to look into the possibility of funding for staff work for the Consortium beyond the current calendar year.

II UPDATE: RECENT AND UPCOMING EVENTS AND ISSUES CONCERNING TESTING

Following a pattern established at previous Consortium conferences, the first full session at the spring 1979 Conference was devoted to sharing of information among representatives of member organizations. Josephine Barresi, of the Council for Exceptional Children, described work on the Council's policy-options project aimed at examining ten areas in education which affect the handicapped in significant ways. One of the topics considered is high school graduation requirements and their effects on handicapped students specifically. She mentioned that a paper describing policy options regarding graduation requirements and their impact on handicapped students is available from the Council (toll free number 800-336-3728). Judah Schwartz, of the Division for Study and Research in Education, of Massachusetts Institute of Technology, reported that the Division is currently undertaking a project concerning the social and political uses of assessment, with support from the Ford Foundation and the National Institute of Education. The project has already organized a panel on assessment in the service of instruction. The hope is that relevant issues will be identified and that papers written on these issues will be circulated and discussed and eventually

published as the first in a series of edited volumes on new perspectives in assessment (hopefully to be published by MIT Press.) Subsequent panels will deal with assessment for accountability, assessment for selection and licensure, and the methodological foundations of existing assessment practice. John Maxwell of the National Council of Teachers of English, described the development of a document entitled "The Essentials of Education," by a gathering of organizations concerned with particular educational subject matter areas. This short paper expresses concern with the current conception of "basics" and affirmatively stakes out what ought to be basic in a larger sense. At present the "Essentials of Education" statement is nearly in final form and as soon as it is available in final form John will send copies to all members of the Consortium, in the hope that other groups may wish to support the statement and perhaps even become signatories to it.

David Seeley, of the Public Education Association(PEA) briefly described the response in New York City to minimum competency testing. He noted that PEA has moved over the past year from a skeptical view of minimum competency testing to one of support. It is not, he went on to say, that PEA is unfamiliar with the problems of competency testing. Nevertheless, PEA has become aware of the important challenge that competency testing poses as to whether we are going to accept yet another scheme for labelling and casting aside those who have failed, or whether we are going to find a way to ensure that all kids do learn the basic skills. David said that he thinks we should accept the positive challenge implicit in minimum competency testing and in this spirit he and others in New York City are organizing a "Campaign for

Competency" to help ensure that kids do learn the basic skills. He mentioned that several groups, including several which originally opposed competency testing have joined in this effort in the spirit that since we have these tests, let's make sure that children do get help in succeeding on them.

Frances Quinto of the National Education Association, described the NEA's freedom of information request to the Federal Trade Commission to obtain a copy of the FTC report on test preparation and coaching. The NEA's interest in the report is twofold. First, if attending costly coaching schools can help students get higher scores on tests which publishers claim are coaching-proof, then this leads to an inequality in education that the NEA is truly concerned about. Second, if this is so, how can we rely on these test results in making crucial decisions about peoples lives? Frances mentioned that the NEA so far has been denied access to the FTC report and data on test coaching schools, but that it will continue its effort to gain access to these materials because it believes the public has a right to know this information. In response to a question Frances said that the FTC has explained it's denial of the NEA request by saying that the data and the draft report on coaching schools are undergoing further review. When queried on the substance of the FTC study, she said she had little information since the study has not yet been released. Nevertheless, she suggested that more detail on the FTC study might be available from Arthur Levine in the panel on test preparation and coaching (see part III). Commenting on Frances' remarks Hayes Mizell wondered aloud whether anyone had established proprietery coaching schools to help students pass

minimum competency tests.

James Ward of the American Federation of Teachers, described an AFT project to help improve teachers' understanding and use of tests. He said that this project is following the pattern of an earlier AFT project on school finance, funded by NIE and designed to bring teachers and policymakers together in a dialogue, and to provide training concerning school finance. The testing project aims over the next six to nine months to develop a training manual concerning standardized tests, and specifically test choice, use and interpretation. The project, funded by NIE, will also encompass a systematic needs assessment of teachers to find out what teachers know,and what they want to know and what they think are the significant issues concerning standardized testing. Once the training manual is developed the project will then organize a series of regional training conferences, attended primarily by teachers, but as was the case in the school finance project, will also be attended by people in policy positions. Jim suggested that hopefully the materials developed will be of sufficient generality that they can be useful to others as well. He suggested too that the AFT hopes to be able to come before the Consortium at future meetings to report significant progress on the project. In response to questions, Jim said that focus of the project would be on most commonly used standardized tests and on minimum competency tests and that it would not focus on teacher -made tests. However, in at least general terms the project would deal with alternatives to standardized tests.

Walt Haney communicated to the group some of Deborah Meier's concerns about this spring's administration of a citywide reading test in New York City. Because of her duties as an elementary school principal Debbie could not attend the first day of the conference, so had asked Walt to communicate her concerns so that interested parties might meet together on Tuesday to talk about the New York reading test. Debbie is concerned about the likely impact on educational policy of the results of the newly revised Metropolitan Reading test, published by the Psychological Corporation and administered to all New York City public school pupils in grades 2- 9 in March, 1979. Specifically she and others are worried that unless a close look is taken at the nature of that test, the results may well surprise a lot of people and influence educational policy in New York in unintended ways. First, on its surface the new test appears to be far more difficult than reading tests previously used, particularly at certain grade levels. Second, the increased test difficulty does not seem to be offset in terms of the number of correct answers required for children to reach 50th percentile levels. Third the new test's grade equivalent tables appear to be designed to produce quite a different grade level score distribution than other tests -- specifically it appears that a higher proportion of children would, on the basis of this test, be labeled as two or more years below and above grade level. Since

New York City relies almost exclusively on grade equivelent scores, despite the publishers warning, this may prove quite misleading. Fourth, data available on the norming of the new test raises questions about its norming, in particular whether under-representation of large urban schools in the norm group may have led to a test of increased difficulty. Because of these concerns, Debbie would like to talk over use of the new test in New York with anyone interested.

Ann Kahn of the National Congress of Parents and Teachers announced the publication by the national PTA of a booklet entitled the Testing Maze by Charlotte Ryan. It represents, she explained, an effort to sensitize PTA members to some of the issues concerning testing in education and to do so in a way that will allow them to delve further into such issues in their own schools and communities. She mentioned that a section in the back of the booklet is a parents' guide through the "testing maze" and raises questions which parents ought to be asking about testing. This section, she said, seems to be particularly useful to parents.* Ann mentioned that the PTA is developing a series of articles on testing for publication in the PTA journal PTA Today (the first of these articles is provided in attachment 1).

---

* The Testing Maze a twenty page booklet by Charlotte Ryan is available for $2.00 from the National PTA, 700 North Rush Street, Chicago, Ill. 60611.

She noted that this journal is distributed to 32,000 local unit presidents, and that the PTA would be continuing to follow testing issues.

Hayes Mizell of the Southeastern Public Education Program mentioned the availability of two publications regarding handicapped students and minimum competency testing: one, published by the National Association of State Directors of Special Education, is somewhat of a survey of minimum competency programs and graduation requirements for the handicapped; the second, a paper prepared by the Education for the Handicapped Policy Project of the Institute for Educational Leadership in Washington, D. C. is more a think piece looking at the relationship between minimum competency programs and the interests of the handicapped, and in particular at relevant federal laws.

Gregory Kelley, managing editor of the Measuring Cup announced the availability of the first two issues of this publication and mentioned that the third issue would be out within the next several weeks. Subscriptions to the periodical are available for $15/year for individuals and $25/year for institutions from:

The Measuring Cup
P. O. Box 22723
Savannah, Georgia 31405

Barbara Lieb-Brilhart of the Speech Communication Association, mentioned that the Elementary and Secondary Education Act, Title II, contains a description of basic skills that includes oral communication

and that there are standards out, distributed by USOE and contributed to by the National Council of Teachers of English, and the Speech Communication Association. These describe what effective oral communication programs look like and emphasizes the need for approaches to assessment. Information on writing standards is also available from the National Council of Teachers of English. The standards for effective oral communications programs are available from:

Speech Communication Association
5205 Leesburg Pike
Falls Church, Va. 22041

Susan Brown of the Mexican American Legal Defense and Education Fund apprised the group of a project of that organization to study alternative methods of law school admissions and said that MALDEF would be glad to share information on this project with anyone interested.

Vito Perrone, summarized activities on the agenda of the North Dakota Study Group on Evaluation for this year. One is completion of what the group is calling a parent's guide to reading tests. Two of the planned five booklets in this series already are available. Standardized Reading Tests: Do They Help or Hurt Your Child and Standardized Reading Tests: What Does That Score Mean? Other booklets are currently being developed. A second area of concern is an effort to find financial support for the Children's Archives (connected with the Prospect School in Bennington, Vt.) which contain longitudinal samples of children's writing and artwork over a fairly extensive period.

A third endeavor involves a collaborative effort to better describe reading and especially how children initially develop reading skills and how they grow into reading. Another major area concerns teachers as researchers and collaborators in research. Minimal competency testing is another concern of the Study Group and it has prepared a new proposal to NIE to study minimal competency testing and in particular, its impact in the classroom, and the politics of competency testing.

Eric Gardner of the National Council on Educational Measurment mentioned the Council's appreciation for all of the problems and issues previously mentioned. He emphasized, however, the complexity of many of these problems and stressed that in our search for solutions we should not overlook the complexity of such problems. They have been with us for a long time, he continued, and surely will not be solved by any single test, policy or law.

James Alatis of Teachers of English to Speakers of Other Languages did not make an oral presentation during the morning session but later did make available a copy of his group's memorandum entitled "Concerns About Statewide Competency Testing". A copy of this memo is provided as attachment 2.

III PANEL ON TEST PREPARATION AND COACHING

The panel on test preparation and coaching was chaired by David Cohen. He explained that despite the hopes of many, no FTC representative would be present since the FTC Report on coaching has not yet been made available. The panelists, then, were only two: Stephen Ivens, director of research at the College Entrance Examination Board and Arthur Levine, an attorney in private practice who was formerly with the FTC.

Steve opened the panel with a presentation regarding the area of test preparation as it relates to the Scholastic Aptitude Test (SAT). He noted that his comments were restricted to the College Board and the SAT. The issue of testing is now in the public domain, according to Steve, and this is for the good insofar as all parties at interest will certainly benefit from discussions such as this one.

Steve began with a discussion of a 1965 statement made by the trustees of the College Board reflecting their concern over the use of special drills for the SAT. The statement was to the effect that intensive drilling (cramming) for the SAT was likely to have little or no effect in raising SAT scores. These intensive drills, Steve said, are what the College Board defined in 1965 as test coaching. The public today defines coaching differently, i.e. as whatever it is that coaching schools do. In Steve's view, this begs the issue of the effect of intensive drill or cramming. This

is the definition the Board has used in its research and in its advice to parents and test-takers. He noted that students' abilities, as measured by the SAT, do change over time. These changes complicate the task of measuring the effects of cramming. The changes, according to Steve, are attributable to two factors: 1) a growth in the abilities being measured, and 2) a lack of perfect measurement precision by the SAT. One of the major problems in studying the effects of preparation, Steve feels, is that students' scores will change - up or down - even in the absence of special preparation. It is a most difficult task to attribute change to any particular factor, Steve said.

The College Board has conducted a number of studies in this area. Steve commented that he was surprised that the area had been studied so little by groups other than the Board and ETS. Nonetheless, while welcoming the FTC to the research arena, Steve said that despite the work of many, a definitive answer to the question of the effects of special preparation remains elusive. He noted that in judging the value of any inquiry on this topic, we must ask a series of questions, relevent both to designing a new study and to reviewing the value of one already conducted. Steve pointed to three studies conducted or funded in the past year by the College Board. The first was a literature review done at ETS by Lewis Pike which Steve feels is the best review, to his knowledge, of the area. This report concluded that there were meaningful effects for the SAT mathematical sections but unresolved

effects on the verbal sections. The second study looked at the use of, and reactions to, a new CEEB booklet, "Taking the SAT." While student reaction was favorable, no score changes were detected as a result of the booklet's use. The third study examined the effectiveness of preparation for the verbal portion of the SAT at eight schools. This study showed a range of -3 to +28 points, with an overall increase of 8 to 9 points.

Steve drew the following conclusions. 1) Students should not expect a big score increase from intensive drilling on sample questions for the SAT. This is, he noted, the same position the Board took in 1965. 2) Instruction in test-taking skills may be beneficial for some students. Steve questioned whether a majority of SAT takers, however, would benefit. 3) A mathmatics review may be beneficial for students not enrolled in a math course at the time the SAT is taken. Steve said the College Board will continue its investigations into this area and welcomes others to do so as well.

Arthur Levine opened with a brief explanation of Freedom of Information Act requests for release of the FTC study which he wrote, as well as for the underlying data for the report. Regarding test coaching, Arthur suggested that the College Board was defining the problem out of existence by saying anything that works isn't coaching. On the FTC study, Arthur said the investigation was originally designed to investigate whether a coaching school could

substantiate its claim of test improvement. According to Arthur, the claim couldn't be proved which left open the question of whether there was a benefit or not. The study followed as an effort to explore these claims, and was publicly announced by the FTC on November 2, 1976. The need then according to Arthur, was for data. The study was limited to commercial test schools that were located through newspaper ads and student interviews. The process was to match up students who were coached with those who were not, and scores were then compared and analyzed. The analysis was concluded in September 1978, according to Arthur, but the report has not yet been released. This was not the ultimate study on the topic, said Arthur, but he expressed puzzlement at the FTC delay in releasing the underlying data which the public could then use to reach its own conclusions.

Arthur said he began his investigation with a belief that coaching was not effective. However, after a review of existing studies, such as the Pike-Evans work of 1972, Arthur came to the view that the issue was open. He then professed distress at statements he attributed to the College Board which stated that coaching had a corrupting effect. Arthur feels that the advantage noted in SAT preparation on math items, coupled with the advantages of "testwiseness" mean the SAT is coachable. He stated that numerous undergraduate institutions which are members of the College Board offer coaching for graduate-level tests and

suggested that such practices represent a contradiction. These practices, he asserted, suggest that these institutions must either acknowledge similar effectiveness for SAT coaching or that they are wasting students' time by offering coaching. Arthur said that St. John University offers coaching for the SAT itself while seemingly adhering to the view in its admissions policy that the SAT is impervious to coaching. This is, he suggested, a particularly blatent example of institutional inconsistency on the issue of coaching.

According to Arthur, the CEEB and ETS have been changing their positions on coaching. He feels that the policy considerations are serious. He says he doesn't endorse coaching schools, but if they are effective, then standardized admissions examinations are invalid and unreliable. He concluded that a major effort at further research is required.

These statements were followed by a lively round of questions and answers from the audience. In response to one query, Steve reiterated his view that the SAT does not measure immutable traits, but the question is what kinds of instruction have what kind of impact over what period of time. Drill on sample test items doesn't seem to work; review of mathematics apparently does work for people not taking a math class; the Board hasn't found anything that affects verbal scores in a similar way. Time on task

seems to be an important variable. Steve also suggested that while a coaching problem is now perhaps being defined into existence, the Board may be as much at fault as anyone by continuing to define coaching as intensive drilling, not as something broader which the public considers to be coaching. Dave Seeley asked whether the preparation techniques which have some effect of increasing SAT test scores have any other educational value as well. Arthur said the FTC report touches on that issue but that he isn't free to discuss that. Steve supposed that there is some value but doesn't know of data regarding the long-term benefit of test preparation. In response to a question from Roger Shuy regarding whether test preparation has any beneficial social or other side-effects that aren't measured on the SAT, Arthur said that there is some literature on anxiety in test taking. He also suggested that the majority of students he worked with while at the FTC did think, at least subjectively, that the preparation reduced their anxiety level. Steve said the PSAT has the same effect. Banesh Hoffman brought the session to a close by observing that in his view if the FTC now does not bring suit against the coaching schools there is a strong inference that the schools are indeed effective (or else the FTC is remiss in failing to protect the public). Banesh also offered a suggestion to the College Board which he felt might solve the problem - namely, if a student could gain a 10-point advantage by paying $50 to take a coaching course, why doesn't the College Board offer for $45 to give the student the extra 10 points without cramming?

IV MINIMUM COMPETENCY TESTING

The first panel on Monday afternoon dealt with minimum competency testing. After giving a brief review of the status of the minimum competency testing movement, Vito Perrone introduced the first three panelists for this session. Corinne Scott of the National Institute of Education briefly described the manner in which NIE became involved in the minimum competency testing controversy. Since NIE does not sponsor programs, the logical way for NIE to address minimum competency testing issues was through research, specifically through research designed to study the impact of such testing on selected student, teacher and school variables. So in July of last year an RFP was issued soliciting such a study. Proposals were received in August. After review by four NIE staff members and four external reviewers, a contract was awarded to National Evaluation Systems (NES) of Amherst, Massachusetts to conduct a study designed to do the following four things:

1. To identify the major types of minimum competency testing programs at the state and local levels,
2. To create a typology which would describe the chief features of these programs,
3. To produce a series of practical guides for educators for use in planning or modifying their own minimum competency testing programs; and
4. To develop the design of a study to assess the impact of minimum competency testing programs on a number of student and school outcomes.

Over a four year period the funding for this study is approximately

$500,000. It is interesting to note, she observed, that the NES is in one way uniquely qualified to do this, because they not only produce minimum competency tests, but now are also evaluating the impact of these tests. They also produce other types of tests. They develop and administer tests for radio operators for the FCC they do certification for architects, self-assessment for physicians, prepare readiness tests for kindergarten children. They also work in preparing basic skills tests, literacy tests, and administer large-scale assessments and do evaluations of educational programs.

Corinne then introduced Bill Gorth, Director of the NES Study of Minimum Competency Testing Programs. Bill described the NES project and how it is organized. Since most of the substance of what Bill said was contained in a project description, we will not summarize his remarks. For the description of the NES Study of Minimum Competency Testing Programs, see attachment 3. Since Bill's answers to several questions from the floor are not covered in the project description, these exchanges are briefly summarized. One person inquired as to when descriptions of minimum competency test programs prepared under the first project task will be available. Bill said that these descriptions will be delivered to NIE in late August or early September. In response to a question concerning the extent to which NES will identify the legislation under which competency testing

programs are being promulgated, Bill said that the project will attempt only to identify sources of authorization, but will not describe or analyze them in detail. Two questions concerned the extent to which minorities' interests would be taken into account in the course of the project. Bill said that their interests would be taken into account indirectly through the project's advisory panel and also through efforts to let a variety of groups know of the project and to solicit their comments and reactions to it. Another question concerned how the states to be visited had been selected. Bill stated that all of the states that have a mandated program will be visited. At present thirty-six states have been scheduled for visits. In addition twenty local school systems which have taken different approaches to minimum competency testing will be visited and their approaches documented. Yet another person asked how NES will deal with the fact that many places do not yet have competency testing programs, but only mandates or plans for them. Bill said that this is a real issue in the NES work, and that all that the project will be able to do is to document the status of minimum competency testing mandates, plans and/or programs as they exist as of about June 30 of this year. Another questioner asked whether there are any plans for having the design of the impact study reviewed by measurement and other experts. Bill said that some such review would be entailed in general review procedures for the project as a whole, but

that he would welcome suggestions for other approaches to such reviews. After Bill finished his formal presentation, one listener asked why the study was designed to produce resource guides on how to implement competency testing programs before carrying out the study to determine whether competency testing has any positive impact. Bill explained that the structure of the program follows from the reality that such programs are already being implemented and so providing information on alternative approaches will be useful even without trying to reach judgments as to the overall value of such programs. In response to another question, he explained further that the project is geared toward implementing competency testing programs, and that this orientation follows from the reality that many states and localities already have mandated that such programs be implemented and so the project is aimed at helping them do so.

Following Bill Gorth's presentation, Hayes Mizell explained his concern for The National Consortium on Testing's goal of promoting public understanding of standardized testing in general and minimum competency testing specifically. In particular he suggested that in many cases, people considering minimum competency testing programs have nowhere to turn to get comments and suggestions on how such programs can be sensibly implemented. Hayes proposed that the Consortium might usefully organize among its members

some means for experts from member organizations to provide feedback and suggestions on local minimum competency testing programs. In this way, Hayes suggested, the Consortium might put itself in the public arena more directly, not to take a position for or against competency testing, but to draw on its expertise and insights to help people in the field who are trying to deal with such testing programs. He suggested that there are probably many people, classroom teachers, legislators and others, who have nowhere to turn to get feedback on particular tests they're considering using, on the strengths and weaknesses of particular tests or on how to use them sensibly.

Following Hayes' presentation, several questions were raised about how Hayes' proposal might be implemented. One person suggested that since it would be hard for the Consortium to serve the needs of many localities or states, it might be possible to write up reviews or case studies of one or two minimum competency testing programs and make these available to others. This led to a question about current circulation of Consortium materials. Walt Haney explained that staff circulars and conference summaries are distributed to member organizations. Also information from the Consortium is currently distributed to a mailing list of 250-300 other people and institutions, including a fair number of state and local education agencies. Another question concerned the extent to which any review panel could represent

the Consortium or even individual member organizations' views on particular approaches to competency testing. Hayes suggested that this would likely not be possible and that any report or view of such a panel might have to be disclaimed as not necessarily representing those of anyone other then individual panel members. But despite such problems, Hayes emphasized, he hoped some means could be devised to make the ostensible expertise represented within the Consortium available to people in the field who are struggling with implementing competency testing programs, usually without much access to technical expertise. Another question concerned the relationship between Hayes' proposal and the NES study of competency testing - which is aimed in part with providing information to states and localities in much the same way as Hayes proposed. Hayes said that in his view the information from the NIE study would come too late and not be of much help anyway. In his view and experience, the NIE study would simply not be useful to people in dealing with specific substantive issues which must be resolved fairly quickly. The discussion ended with the observation that Hayes' proposal would be discussed further in a small group session on Tuesday morning.

V PROGRAM EVALUATION AND TESTING

The first presentation in this session was by Daniel Stufflebeam of Western Michigan University, who described development of standards for educational evaluation. Recounting an anecdote concerning application of industrial standards, Dan suggested that standards such as the educational evaluation standards probably ought not be taken too seriously -- because there will be all sorts of problems in actually trying to apply them. Nevertheless, what he wanted to accomplish at the Consortium meeting was two things: 1) To tell people about the evaluation standards, specifically the fourth draft of these standards -- which hopefully will be the second to last version of them and; 2) To solicit peoples reactions, criticisms and comments concerning them. He also invited people to comment on or criticize the standards either by attending hearings on the standards in Washington, D. C. on May 2, or by writing directly to Dan at Western Michigan University. Anyone wishing to obtain a copy of the draft Standards for Educational Evaluation can get one from the American Psychological Association for $3.00.

Dan then distributed a handout consisting of the Table-of-Contents to the Standards, a list of members of the joint committee which developed the standards, a brief summary of the Standards, a detailed example of one standard, an index to the Standards, and a schedule of the project for developing the Standards (See attachment 4.) He briefly explained that there are twenty-nine standards cutting across four broad attributes which any good evaluation should encompass. These have to do with accuracy, utility, propriety, and feasibility. Since Dan's handout, a copy of which is provided in attachment 4 gives many of the details of Dan's talk we do not summarize it in detail here.

After describing the development of the draft Standards, Dan explained that the joint standards committee views the Standards not as a static document, but instead as something which should be revised and adapted on the basis of experience. He closed his presentation by reading a letter commenting on the draft Standards, raising questions about their impact and means of enforcement. He emphasized that those responsible for development of the standards welcome more such critical scrutiny and commentary on the Standards.

In the second presentation concerning program evaluation and testing, Walt Haney described the USOE's proposed system for the evaluation of Title I projects. He began by noting that there has already been considerable controversy over the system, and that originally it had been hoped to have both proponents and opponents of the system attend the conference to describe their perspectives on the Title I evaluation system. However, because of the tight schedule for the Consortium conference there was simply not sufficient time available to give such extensive treatment to this topic. As a result, Walt explained he would try to describe at least fairly fairly, the rationale and reasons behind development of the system and also the sources of opposition to it. After briefly describing the history of Title I evaluation, Walt noted that the proposed USOE system for evaluating Title I has been developed as a direct result of a congressional mandate for evaluating Title I programs. Specifically, the 1974 Education Amendments, directed the Commissioner of Education to

"provide state education agencies with models for evaluation of all programs conducted under Title I." The idea was that comparable results from such models could be aggregated across states to yield results on the effectiveness of Title I nationwide. There are, Walt went on, essentially three different evaluation models which have been proposed as part of the Title I evaluation system:

A - Norm-Referenced Model

B - Comparison Group Model

C - Special Regression Model

Each of the models is aimed at estimating the effectiveness of Title I programs in terms subtracting some estimate of what participating children's test scores would have been without Title I from the actual observed score which children actually do achieve at tne end of a Title I program. The three models differ mainly in terms of how they go about estimating what childrens test scores would have been without Title I. According to USOE proposed regulations published in the Federal Register, Feb. 7, 1979, these models are applicable for all LEAs funded under the Basic Grants program of Title I which provide instruction in reading, language arts and math, in grades 2-12. The proposed regulations describe the three models as follows:

> The Norm-Referenced Model uses the strategy of pre-and post-testing the Title I children, and comparing their performance to the table of norms accompanying the test. In this model, expected performance is estimated by assuming that, without Title I services, the Title I

> children would maintain their percentile status relative to a national or local norm group from pre-test to post-test...
>
> The Comparison Group Model is the classical experimental design in which the performance of Title I children is compared to that of another group...
>
> In the Special Regression Model, a group of children is divided into Title I and comparison groups based on a pretest cutoff score. Title I services are provided to children scoring below the cutoff. Children scoring above the cutoff are the comparison group for the evaluation. Expected performance is estimated from the pretest and post-test scores of the comparison group by applying a statistical procedure known as the regression model. This model, when properly applied, will yield an estimate of expected performance that takes into account differences between the two groups that are not the result of Title I services.*

The proposed regulations, Walt continued, specify that in order to make results of evaluations comparable, procedures have been developed for summarizing data in terms of a common reporting metric. The proposed regulations do not specify what this metric is, but it seems clear, according to USOE materials that this will be the normal curve equivalent (NCE). The NCE would appear, given the attention it is arousing, to be something new. Already new versions of the California Achievement Test, the Gates-MacGinitie Reading Test, the Iowa Achievement Test, the Metropolitan Achievement Test and the SRA Achievement Series are providing

---

* Federal Register - 44:27, Wednesday February 7, 1979, 7916.

norms tables expressed in terms of NCEs. However, the NCE is really nothing more than a standard deviation divided by 21.06. Thus 98% of a normal distribution fall between 1 and 99 NCEs. This makes the range of NCEs approximately equal to the range of percentile scores. However, unlike percentile scores, according to the rationale behind the new evaluation system, the NCE constitutes an equal interval scale, and thus, unlike percentile scores provides a suitable basis for aggregation of results across evaluations. Having explained the rationale for the USOE proposed models, Walt went on to review criticisms of the new evaluation systems. He noted that for each of the last 3 years there have been a dozen or more papers dealing with the new Title I evaluation system, or at least with particular aspects of it, at the annual AERA meetings. Prominent critics of the Title I evaluation system include Richard Jaeger, Robert Linn and David Wiley. Linn has recently written that the USOE proposed regulations will not adequately serve the intents of Congress. Rather, in his view, "They will instead yield potentially misleading results and will stifle the collection of information of much greater utility."

Criticisms of the USOE proposed evaluation system seem to have fallen roughly into three areas:

1 - specific criticisms of each of the three models;

2 - criticisms of the comparability of results from two or more of the models;

3 - criticisms of the utility of the data to be derived and arguments favoring the collection of alternative data.

Since time is short, Walt continued, he would skip over the first

two sorts of criticisms and focus on the third type. The first general criticism concerns the premise that implementation of the three models will yield valid and comparable results. However, little empirical evidence is available to support this premise and indeed some empirical work contradicts the premise that the alternative models will yield comparable results. Second is that a broader array of evaluation approaches is needed in order to meet the diverse information needs of people involved and concerned with Title I at federal, state and local levels. Third is that standardized achievement test results provide only one part of the evaluative information needed to provide sensible program evaluation. With respect to this point, Walt suggested that the proposed Title I evaluation system represents a considerable contrast with many of the issues raised in the evaluation standards discussed by Dan Stufflebeam. Fourth, is that the major rationale behind the Title I system seems to be the aggregation of results across localities and states so as to derive a national picture of Title I effectiveness. However, if this is a major goal, critics suggest it probably could be accomplished far more efficiently and accurately by designing a systematic sampling approach to evaluating Title I projects. The fifth general concern is probably the most important, and Walt suggested, has been well expressed by Robert Linn. So he quoted Linn's remarks criticizing the proposed system. Specifically Linn criticized the very heavy reliance of the system on norm-referenced achievement tests:

> By their very nature, such tests are rather global. They cannot be expected to be as sensitive to specific project effects as measures that are keyed to the emphases of the project. I would not argue that project specific measures should replace the broader survey measurement provided by traditional norm-referenced tests. Narrow measures that are highly tailored to project goals are insufficient by themselves because they do not provide information about more global achievement. On the other hand, the global information provided by norm-referenced tests is insufficient by itself because the information is not specific enough to have much use in improving the program and lacks the sensitivity to detect small effects in particular content areas.*

Following this presentation, questions and discussion concerned both the evaluation standards and the USOE proposed Title I evaluation system. One person asked whether the Standards project had considered treating the issue of selecting outcome measures. Dan Stufflebeam recounted that the project decided not to emphasize testing because it did not want to get trapped into the appearance of equating testing with evaluation. Nevertheless, he pointed out that the Standards do deal with many of the issues raised, for example, with respect to Title I evaluation, but at a higher level of generality than they often are considered. Eric Gardner added that the specific issues of test use in evaluation is being considered by the committee formed to review the joint test standards, and that this committee has given attention to the specific issue of test use. Several people pointed attention to the contrast between the USOE Title I evaluation

*Robert Linn, "Comments on the Proposed Regulations and Models for Title I Evaluation," unpublished paper, University of Illinois at Urbana-Champaign, p. 13.

system and the draft evaluation Standards. One observer expressed dismay that the evaluation Standards are not yet publicly available so that they could be used to illuminate debates over issues such as Title I evaluation. To this comment Dan expressed sympathy but noted that parties to the development of the Standards naturally do not want them released until they have been thoroughly reviewed and agreed upon. Another commentator said that the standards seem reasonable in certain respects, but seem rather fatuous in that they fail to direct attention to the substance of education and to students. In response to another comment suggesting that the evaluation standards are more oriented toward evaluators than to potential users of evaluation results, Dan suggested that it may be necessary to produce a number of derivative documents aimed at other audiences. Another observer suggested that despite the criticisms of the USOE proposed models for evaluating Title I programs, they do seem to represent a significant improvement over previously prevailing practices in Title I evaluation. Banesh Hoffman, commenting not just on issues raised in presentations concerning program evaluation and testing, but also on the presentations concerning minimum competency testing, suggested that evaluation ultimately may be more art than science. To make his point he asked people to imagine that a National Institute of Humanities rather than the National Institute of Education was organizing an evaluation. How, he asked, would such an NIH go about evaluating literature or art? Do you suppose that it would develop a set of standards for such evaluation or would go about evaluating books or art in the same manner as the NIE seems to be going about evaluating minimum competency testing?

VI. VIEWING AND DISCUSSION OF PUBLIC INFORMATION FILMS ON TESTING

On Monday evening, a small group of interested people gathered to view and discuss two films designed to educate the public about testing. The two films were selected for viewing because they were both of recent vintage and because they represented very different perspectives on testing. The two films, and their availability are described as follows:

Let's Talk about the SAT

16 mm sound and color film -- 14 minutes

Presents a candid picture of the Scholastic Aptitude Test (SAT) as seen through the eyes of those most closely associated with the test -- high school and college students, counselors, and college admissions directors. It is designed to inform students about the test and to alleviate the anxieties and apprehensions they might have.

This film is available on loan for a $5 service charge (to cover handling, maintenance, and postage) or for purchase at $120 per print. If you wish to buy a print following initial screening of the film, the $5 service charge will be deducted from the purchase price. A free discussion guide will accompany each order.

College Board Film Library
470 Park Avenue So.
New York, N. Y. 10016
212/684-5910

Testing -- None of the Above

a University of Wisconsin-Extension film
color, optical sound, 16 mm, 33 minutes

Summary: An unusual turn of events enables a school counselor to see her work from a new point of view. She looks at several aspects of the use of standardized tests: the interpretation of standardized test scores to label a child an overachiever, attitudes toward knowledge of I.Q. scores, the effect of test scores on teacher expectations, grade level norms, the scarcity concept of education, and the validity of test items. The counselor discovers points of view she had been ignoring. Her new realizations allow her to return to work renewed.

To rent copies of Testing--None of the Above
write or phone:

Bureau of Audio-Visual Instruction
University of Wisconsin - Extension
1327 University Avenue
P. O. Box 2093
Madison, Wi. 53701
phone: (608) 262-1644
catalogue number: 1817

Rates: One week rental - Mon.-Fri. $11.25 in Wisconsin, $21.25 outside of Wisconsin. Purchase price, $400.

VII TOPICAL SESSIONS

On Tuesday morning three topical sessions were held dealing with:

- Information on Testing for Parents and the Public
- Testing and Program Evaluation
- Minimum Competency Testing

Due to lack of time later in the day, the discussions and recommendations made in these three small group discussions were never summarized before the full conference. Since no summaries of these sessions were prepared, we recount here only specific suggestions raised in these sessions for future Consortium activities and meetings.

Three specific suggestions were made in the session on public education concerning testing. First, several people suggested that at a future meeting the NCT ought to give attention to the relationship between school curriculum and school testing, and to the effects of testing on curriculum and on teaching. Second, Al Dyson suggested that it would be very useful for the Consortium to prepare an analysis of the legal issues involved in parents gaining access to and acquiring copies of tests which their children are required to take. Third, it was suggested that the Consortium might usefully prepare a one-page list of questions which parents ought to ask about school testing programs in their own communities. With respect to this third suggestion it was noted that one version

of such a question list is already available in the PTA publication the Testing Maze.

Discussions of testing and program evaluation resulted in two specific suggestions. First, it was suggested that the Consortium should attempt to keep abreast of developments concerning the educational evaluation standards. Second, it was recommended that the Consortium give attention to testing policies in the state of New Jersey where student test results are apparently being considered as a means of assessing teacher performance and also as a basis for funds allocation.

In the session on minimum competency testing, support was expressed for Hayes Mizell's proposal, but several questions raised about how it might be implemented. Also, it was suggested that the Consortium might usefully keep track of state legislation concerning testing.

VIII OPEN VERSUS SECURE TESTING

Vito Perrone introduced this panel and chaired the subsequent discussion. He mentioned several legislative initiatives regarding open testing and framed several key questions in the area: Will open testing subvert the technical qualities of existing tests? Will it force the costs of tests to an unreasonable level? Or will it bring greater fairness to testing and an improvement in the quality of tests? Walt Haney spoke briefly of the work at the Huron Institute on this issue. He mentioned the draft paper circulated to some Confortium members which attempts to sort out some of the issues involved. Andy Strenio, the author of the draft, explained the process of preparing a staff circular on open versus secure testing and thanked those who had contributed material and comments on the draft.

The four panel members were: William Corbett, principal of the James Russell Lowell School of Watertown, Massachusetts, and prime mover behind a Massachusetts regulation which guarantees students and parents access to the Massachusetts basic skills tests after they've been administered in the state; Alice Irby, Vice President and Washington representative of the Educational Testing Service (ETS); Stephen Solomon of the New York Public Interest Research Group (NYPIRG) which has been instrumental in advocating "truth-in-testing" legislation in New York; and Winsor Lott, Chief of the Bureau of Elementary and Secondary Testing for the New York State Department of Education.

The first panelist, Bill Corbett, opened his talk by sharing with the audience one of the first "competency tests" ever administered in this country. The test was given to an eighth grade class in Watertown, Mass., 104 years ago as a requirement for admission to Watertown High School. One of the questions was: "What is the product of the square root of 819.75 times the cube root of .001906624?" Bill made copies of the rest of the test available to the audience after his talk.

Bill said he had been concerned with the processes and secrecy of testing during most of his 30 years in education, but that this concern became submerged with other things he had to do. Norm-referenced standardized testing of little children was most irksome to Bill, especially with regard to the bell-shaped curve. Bill said that he was never reconciled with the notion that, for second-grade children, for example, half of them would fall above and the other half below the mean before the test was even administered. According to Bill, the durability of the bell-shaped curve is partly attributable to the training in psychometrics which teachers and principals encounter in their preparation. Bill thinks that statistically the procedure is beautiful and valuable, but it has no place in the evaluation of little children.

Machines are subject to mechanical error and are fed by attendants subject to human error. The numbers become the focus of attention, according to Bill, since they are precise while the shortcomings of the tests are overlooked. Few parents until now have asked to see the test questions or corrected answers because the test companies have made secrecy a way of life with educators. Bill said educators have

become partners in test secrecy and that they don't like it. Massachusetts was planning on becoming the 37th state to adopt competency testing in June 1978. At the same time, Bill's son took the MCAT for medical school admission, and a human error caused 90% of MCAT test-takers to receive scores lower than they earned. While the net result was rectified, it had, in Bill's opinion, latent "Three-Mile Island" possibilities. In any event, it moved Bill to appear before the Massachusetts Board of Education as a private citizen to request a form of openness in the tests: one that would allow parents and students to see the tests used in basic skills assessment along with the corrected answers. The Board adopted such a policy and it was incorporated into the regulations on January 23, 1979. Bill hopes that similar regulations will be adopted nationwide.

Bill said the process of implementing open testing will require several steps: 1) The ordering agent must remember to request the optically-scanned answer sheets be returned, a service offered by the major testing companies in their catalogs, and 2) A counselor can then hand-score an optically-scanned answer sheet in a few minutes. The test and scored answer sheet may then be shared with a parent or student upon request. Such restricted sharing, according to Bill, would avoid the pitfalls of copyright and security problems.

Bill concluded that this Massachusetts policy is quite possibly an intermediate one. Implementing the policy may prove difficult, he suggested, but this just tends to prove the statement of Thomas Carlyle that "Every noble work is at first impossible." Bill said that should this Massachusetts policy prove to be the first step in the process of

of breaking what one gentleman has described as the tyranny of testing, he wanted to note that he had suggested the policy only after reading many papers on the subject, many by NCT members, plus direct observation of the effects of testing upon children with whom he had worked over the last three decades.

The second panelist on Tuesday morning was Alice Irby of ETS. Alice's remarks were based upon the text of a paper she co-authored with John Fremer, also of ETS, entitled, "Why Should Some Tests Be Secure?" A copy of the paper is appended to this report as Attachment 5. Alice introduced John Fremer, the Director of Test Development for Higher Education and School Programs at ETS, and explained that he was available to answer any technical questions she might be unfamiliar with.

The third panelist was Stephen Solomon of NYPIRG. Steve began by explaining that in his view truth-in-testing legislation is primarily designed to provide more information about standardized tests and encourage more public discussion about the proper use of tests. Specifically, the proposal before the N.Y. legislature calls for the publication of test questions and answers on tests used for postsecondary and professional school admissions, thirty days after the results of the tests are released. This year's N.Y. proposal has been revised in light of past criticisms,such as worries about equating or pre-testing questions (these would be immune from disclosure in the current bill). As far as cost considerations, according to Steve, it isn't that expensive a process to develop new questions; the cost is generally less than 10% of the total cost for any test. The basic idea, according

to Steve, is that students have the right to know what they're being evaluated on. Even if costs were very high, the benefits of upholding this principle might be worth it anyway. Steve said that if costs of open testing are spread across all test-takers, the cost per person would be small. According to Steve another aspect of "truth-in-testing" legislation is a requirement that all studies concerning the tests be made public by being filed with the state education office. This would provide greater accessibility to information about testing by those concerned with the subject. Also, the bill would require more information be made available to test-takers prior to testing. According to Steve, his experiences with NYPIRG and its 125,000 student members and others, students are not satisfied with the information they are currently receiving concerning secure tests. Steve said students, as paying consumers of the tests, are just as entitled to detailed information as sponsoring bodies or institutions of higher education.

Steve feels that the impetus for the open testing movement comes from a new source: not from those traditionally concerned with testing, i.e., the professionals in the field, but rather from the laypeople, the students, those who are not normally represented at professional gatherings. Steve said NYPIRG was especially gratified, in this regard, to be able to join the Consortium and express its view. Nonetheless, students as test-takers are seeking a greater say than they've been accorded in the past. In addition to the New York bill, he noted that truth-in-testing proposals have been made in several states

(including Texas, Maryland, Colorado, and Indiana) and may be re-introduced before long in Congress. Steve feels that many areas of investigation regarding the merits of open testing haven't fully been explored in the past but are gaining attention.

In addition, Steve thinks it would be useful for representatives of the test industry, students, consumers and educational groups to sit down together and try to work out viable schemes for open testing. Steve suggested that compromises may exist that would make open testing desirable for all concerned groups. He sees NYPIRG's concern as encouraging the proper use of tests and thinks a dialogue and enactment of open testing will contribute to that end. Steve thinks the "mystique of testing" comes from a fear of the SAT and other secure tests and not from a thorough understanding of what they are all about. More openness in testing would be a good thing, Steve said, and could occur while protecting the psychometric validity of tests. In addition, according to Steve, open testing would insure that all students have access to test questions, not just those who can afford to pay for expensive coaching courses. As far as the benefits of coaching -- are there any? If the time and money spent by students on SAT preparation is any indication, one group that feels there are benefits is students.

Steve closed by asserting that truth-in-testing legislation will promote wider understanding of the strengths and weaknesses of standardized tests among educators, legislators, other professionals, and the public at large. He said he'd like to see more discussion between advocates of open testing and the testing industry so that compromises could be made that would eliminate the need for legisla-

tive initiatives.

The fourth and final panelist to promote open testing for the morning session was Winsor Lott. Winsor said that the NY State Education Department publishes mainly open tests. According to Winsor, the Department this school year will develop and publish 36 high school regents examinations, 16 other high school achievement tests, and 10 competency tests, for a total of 62 tests that are completely open. Some 8 other tests that are administered are semi-open; that is, they're available for inspection in school buildings throughout the state. Nonetheless, despite being in favor of open tests and agreeing that students should have an opportunity to go over the tests they've taken and to look at the right answers and question the key, Winsor said the NY Department has gone on record as opposed to the La Valle bill. Nevertheless, he suggested that this year's bill is a great improvement over the version that was discussed last year.

According to Winsor, the Department can produce better secure tests than open tests and at a lesser cost. Past NY experience led him to this conclusion. He feels repeated use of test questions allows for greater refinement of tests over the years. As far as cost considerations go, he said the La Valle bill will lead to a reduction in the number of test administrations within NY and at a greater fee, and that consequently NY candidates for admission to college will be put at a disadvantage. He also suggested that the use of open tests in NY scholarship programs would lead to the greatest cost effect of such testing being placed upon those who have the least ability to pay.

However, according to Winsor, there are other ways to accomplish the goal of openness. Some of these alternatives include making forms available for research purposes which wouldn't have adverse cost implications. Winsor also thought test publishers could take a leaf from the NY Civil Service exams which are totally secure and yet everyone who takes the test gets an opportunity to sit down and go over the test and examine the key and see what they did wrong and challenge the questions. Winsor suggested the same procedure could be followed by the publishers of the national tests that are the object of the La Valle bill. For example, locations could be established throughout the country where students would be given an opportunity to review the questions and the key. Winsor concluded by remarking that now that he's said these things he's going to sit down and get out of the crossfire.

A brief period for questions and comments followed. Paul Jacobs said that the first 5 reasons Alice Irby gave for keeping tests secure were all economic, i.e. that it's less expensive to do things as they are now. Paul said that may be so, but alternatives haven't really been explored. The sixth reason, according to Paul, was a fear of greater teaching to the test. Paul said that if the SAT is truly a test of developed reasoning ability, and if students were taught in terms of a bank of released test questions then that would be wonderful; schools should be teaching reasoning ability. Reed Edwards of Congressman George Miller's office observed that in his university about 75% of the exams were put on open file and that

those files helped students to better understand the types of questions and style of testing that professors were using. Steve Solomon then asked what percentage of questions are re-used on the SAT, feeling that if only a small proportion are currently re-used, an end to test secrecy wouldn't give that much away to begin with. John Fremer responded by objecting to the use of a "Three-Mile Island" analogy for testing and noted that test security had allowed the rapid development of a substitute SAT so people didn't have to go to the area around Harrisburg to take the SAT. He went on to say that 100% of good questions were re-used. Homer Elseroad suggested that the NCT put together a group that would solve the question of costs given various assumptions. Then we would be free to discuss the real questions and not be carried away by vague talk about costs. Walt Haney added that when the Huron study began one specific question that was put to every member of the Consortium and numerous groups was "do you know of any information on the cost of developing standardized secure tests." Judah Schwartz said that one reason for testing was to evaluate the quality of teaching and of schools and that this can't really be done if some of the answers which determine what is or is not taught are not available. Ed Taylor suggested that tests are major instruments of public policy and yet are not available to public inspection. He said you need to look at the entire test to see what goes into the selection process and what prejudices may enter in one way or another. Debbie Meier said that it

would be helpful to know what items students do better or worse on in order to facilitate the teaching process, and that this is related to the idea of open testing. She also thinks it's important to be able to look at your own answers and see if you were close or not close and whether it was for a good or bad reason. Bill Corbett said that while the discussion today focused on upper educational levels, the vast majority of tests are given to little children and there was no one present at the meeting to represent their interests. He suggested that at some point in the future the NCT should look at the effects of testing on little children. John Weiss said he thinks that open testing would lead to a more open evaluation of testing. In his view, the process of improving certain tests and discarding others that are not desirable, of reforming our evaluation instruments, would be aided by adoption of open testing procedures. Marjorie Martus said that, if there is to be a more informed discussion of tests, more studies of tests conducted independent of the test-makers are needed. For that to happen, in her view, it is urgent for more information to be made available for inspection. Alice Irby responded that access for purposes of research is already available for independent researchers in the field. Al Dyson said that if you do a study of bias in test questions but are forbidden to publish the questions themselves, you're in a quandry: if you publish the study without the items, people won't know what your findings are all about.

IX PUBLICLY AVAILABLE TEST ITEM BANKS

Jerrold Zacharias of the Education Development Center delivered a talk proposing development of test item banks which would be open to the public. The talk was introduced by Vito Perrone. Jerrold began with a reference to the Three Mile Island incident as an analogy for testing. He said that if he wanted nuclear power, and he does, he wouldn't hire Babcock & Wilcox to do it. He believes the public needs tests, that they can be made, and that he wouldn't hire Babcock & Wilcox to do it. He continued by saying that for all types of tests that are known to people, the most you can say is one of three things: great, OK, and something's wrong. As soon as you say that you are going to make tests that put people in a numerical order you have already exhibited a degree of intellectual ignorance that should bar you from having anything to do with that operation. If you find from a test that something's wrong, you need to know how to get at whatever's gone wrong. Jerrold thinks that an end to formal secrecy is not enough; what we really need, in his view, is ready and easy availability. Tests should be printed in an inexpensive way -- newsprint -- and easily distributed. He wants general availability of the tests, not only after researchers promise to keep the materials they are shown secret. He sees the creation of test item banks, and he stressed the plural, as the direction in which we ought to move. Everyone would have easy access to test items, criticisms of items and commentary on the test items. Further,

turf would be abolished -- since fields would not be restricted to experts in that field alone, what he calls a resonant discussion could then result. Jerrold observed that testing is to some extent against the democratic grain in the sense that there wasn't a popular election to see who would sculpt the Michelangelo David; they just got Michelangelo. He had carved before and there was access to high quality commentary. If you could include commentary by all minority groups on test items, you'd be better off.

Now, how do you do this? Jerrold said his image is quite simple: like a Sears Roebuck Catalog. You put together a big catalog with little supplements, all printed inexpensively. You need a distribution system as well, one that can be of aid to teachers in their jobs. This can be done from the ground up, i.e. letting the teachers and principals have a major say in the selection of items and then taking advantage of bulk rates and so forth. Computers can handle inventory and commentary on the inventory for an entire bank. Then you could "order up" tests that do certain things: equivalent questions which don't give any cultural advantages. Then the computerized arrangement could provide photo-offset copy to the schools.

Jerrold feels that banks could be started along these lines. He feels discussion is needed among people in small interest groups to move the process along in a positive manner. For those interested in reading more on the subject, Jerrold asks that they write to him and he would be happy to provide more material.

In response to a question, Jerrold elaborated that a teacher could specify what kinds of questions with what sort of response patterns are desired and the computer could handle that, that the computer is wonderful at handling the busy work for us. He said that a certain amount of time would be required to reach an equilibrium as test items are sorted out for value, ambiquity, and the like.

X OPEN COMPETENCE TESTING

The final speaker of the afternoon was David McClelland who addressed the topic of open competence testing. He began by noting that the kind of testing he's most interested in doesn't involve the issue of security. David contrasted two types of tests one of which is wide open: the driver's test. There are driver's schools and no one complains of fairness. We want to know whether the person can drive. There are other tests like this -- typing tests or throwing baskets from the foul line and so forth. This kind of test differs from, say, the SAT. David says the driver's test is a job competency test, i.e. criterion sampling where you're sampling the behavior itself and not some indicator of it. The problem with competency indicators, according to David, is that they get you into what is called norm-referenced testing; since you're measuring an indicator you want to sort people out into better and worse. But the connection between the indicator and the competency itself may be very loose. SAT scores, for instance, don't seem to be indicators of any competencies that most people are interested in, except the competence of taking tests. David said it is astonishing to him how little, at least within certain ranges, these indicators point to real life competencies.

If you're measuring the skill itself, you don't run into fairness problems -- people accept performance differences there. Moving on, David said that a second distinction among these two types of tests

is that a driver's test is an operant rather than a respondent test. That is, you have to make a response rather than simply choosing among alternatives that are presented to you. The problem in choosing among alternatives is that you can cheat more readily than when you have to make responses entirely upon your own. That's the reason security is much more of a problem with respondent tests; that is it is technically possible to memorize a sequence of answers without knowing a skill at all. A third difference is the purpose. For the driver's test, the goal is to pass everyone, to get them up to the standard. Hence there is a collaborative relationship between the tester and the test-taker -- the more open the better, because the real goal is more to help people do better. Not so with the SAT, according to David, because the primary purpose there is to screen people out since everyone can't be admitted to college.

David said his experience has been predominantly with the "Operant Job Competency" type of test. The work has involved extending this approach beyond mechanical skills to evaluation of occupational performance. In doing so, he attempts to: 1) identify the competency desired, and 2) then invent an operant type test to see if the person displays that job competency in a test or interview situation. David says it's important to develop accurate coding systems that are reliable and objective or otherwise bias may enter into the scoring. According to David, this process is unlike essay grading where criteria are often not made explicit.

Finally, a cross-validation technique is employed to see if superior people do in fact score higher on the test. If so, then an explicit coding system exists around which teaching can be designed. He noted that his work often involves teaching rather than selection. According to David, once you make very explicit some test characteristic that you want to be able to improve, you can design your teaching around it. It is a kind of teaching to the test, but you are teaching people to think a certain way. David then discussed some of the work that he has done in his career along these lines for various groups, including the navy.

David thinks that selection isn't always the answer that it's sometimes thought to be because the competencies needed at different levels of organizations are somewhat different. So, if you screen some people out for characterisitcs that are needed for the first level, you may as well be screening out people who have characteristics needed for the second, third or fourth level. So, it's probably more important to train people for particular jobs they're selected for than to rely upon screening on the basis of some generalbility test. David talked of various executive motive profiles and caring motive profiles as being indicators of competency for certain jobs such as alcoholism counseling. In addition, David mentioned the programmed case test where the test-taker is given some information about a person and is then asked to predict that person's behavior given a set

of choices, The better counselors get better at this as the test goes along because they develop an empathy for the "subject" based on their experience and understanding. So, summing up the relationships of four sorts of testing, David drew a 2 x 2 chart:

	Indirect COMPETENCY INDICATOR	Direct JOB COMPETENCY
OPERANT	Executive Motive Profile Caring Motive Profile	Driver's Test
RESPONDENT	SAT	Programmed Case

David said that if you learn enough about the programmed case method you are teaching the skill itself, so it's like the driver's test in that sense. He described what he calls a diagnostic listening test which he suggested might have to be kept secure because it focused on fairly few attributes in evaluating performance which could be easily memorized, and possibly faked.

David said tests represent a two-edged sword: they were originally designed to fight the Boston Brahmins who were putting all their relatives into civil service jobs. The tests were used to make it fairer for the poor who didn't have connections to get the good jobs, and we tend to forget that. Also, he suggested, it is difficult in any system of assessment to prevent people who are strongly self-interested from gaining unfair advantage. In this respect, he cited experiences with the Massachusetts Civil Service Commission. You have to watch out for influence creeping back into any system of assessment and selection,

David cautioned. Nevertheless, he said that he is not opposed to the NY bill just discussed. In conclusion, David observed that in general you can sidestep the security issue by dealing in operant measures of job competency or, if necessary, in respondent measures of job competency. Furthermore, he has found that operant measures are more sensitive indicators of educational effects. David said he would like to see the bottom left cell of his two-by-two table disappear from testing, that he doesn't like the pre-coded items and thinks they became used primarily because they're cheap. David mentioned that he has a book coming out that shows the sensitivity of educational value to school quality, as measured by operant rather than respondent tests. The ideal in David's view would be to design tests to fit the competencies and not just to stay with old tests that have become traditional. We need to start over again, David thinks, to work up tests that are appropriate for the situation, and he feels that such tests will be largely operant.

ATTACHMENT I

The standardized test contributes to the furor. PTA Today March 1979, pp. 12-13.

This attachment was removed prior to submission to the ERIC Document Reproduction Service.

ATTACHMENT 2

Memorandum from TESOL on concerns about statewide competency testing. TESOL Quarterly V13, N1, March 1979.

This attachment was removed prior to submission to the ERIC Document Reproduction Service.

For further information contact:

James E. Alatic, Exec. Sec.
TESOL (Teacher of English to Speakers
of Other Languages
455 Nevils Building
Georgetown University
Washington, D.C 20057

Attachment 3

A STUDY OF MINIMUM COMPETENCY TESTING PROGRAMS

INTRODUCTION

In December 1978, the Office of Testing, Assessment and Evaluation of the National Institute of Education (NIE) initiated a two-phase, four-year study of minimum competency testing programs across the country. The study, being conducted by National Evaluation Systems, Inc. (NES) of Amherst, Massachusetts, is designed to review and analyze the minimum competency testing programs existing in over 30 states and numerous local school districts and to plan an impact study of a sample of these programs. Phase I of the study will extend for the period of one year and will focus on four major tasks: (1) the identification of the major types of minimum competency testing programs currently in place, (2) the creation of a typology describing the chief features of these programs, (3) the production of a series of practical guides for educators to assist in planning or modifying their own minimum competency programs, and (4) the development of a design for a study to assess the impact of minimum competency programs on a number of student and school outcomes. Upon approval of the plan for the impact study, it will be implemented in Phase II of the study and will extend over the following three years.

Background and Context

Minimum competency testing is perhaps the most visible and controversial issue in education today. Support for programs designed to assess whether students have acquired some specified set of skills to a predetermined minimum level is due in large measure to a growing public concern about the condition of the American educational system. Indicators such as the successive 14-year decline of scores on the Scholastic Aptitude Test and statistics that show that one in eight high school graduates cannot read well enough to understand a simple traffic sign have led many people across the country to question the quality of American education.

The breadth and depth of the trend toward minimum competency testing and the rapidity of the growth of the movement have outstripped efforts to determine what these programs in fact achieve. Until now, detailed, reflective information about the impact of minimum competency testing has been unavailable, and educators continue to argue the merits of the programs. Yet, despite this absence of information, educational decision-makers are being charged with designing or refining minimum competency programs in their respective states and local districts. It is in response to the pressing need for information about the outcomes of minimum competency testing and for supporting information about alternative program strategies that NIE has initiated the current study.

Purpose of the Study

The NIE Minimum Competency Study is a research and dissemination effort designed to support state and local decision-makers. The overall purpose is to assess, inform, and evaluate—not to advocate—with the goal of providing an open, objective, and fair description of the minimum competency testing movement. Aspects of this general purpose will be addressed by each phase of the study.

PHASE I. The primary purpose of Phase I is to collect and disseminate information about minimum competency testing programs. Specifically, the focus will be on identifying the types of information that would be most useful to state and local policy-makers, to describe the current state-of-the-art, and to disseminate resource guides that will assist in making informed choices about minimum competency testing.

PHASE II. The chief purpose of Phase II will be to implement an impact study of minimum competency testing programs. Ultimately, the study should generate important indicators of the impact of the movement in order that future decisions in this area may be made on a sound and reliable basis. Information collected during Phase I will ensure that the impact study to be conducted in Phase II will be valid, feasible, and responsive to the variety of existing programs.

Design of Phase 1

TASK 1: Develop Program Descriptions. As one of the primary information collection activities, visits will be made to each of the more than 30 states and a sample of 20 local school districts with existing minimum competency testing programs. Meetings will be held with directors and staff of assessment programs, members of state boards of education and legislatures, lay public advisory committees, superintendents, principals, and local testing coordinators.

Discussion at the meetings will focus on solicitation of suggestions for the design of the impact study plan and topic areas for the program guides as well as detailed information on the respective programs.

Program details of interest will include:

1) purpose and historical origins;
2) governing legislation, rules and regulations;
3) skill areas and competencies assessed;
4) procedures used to identify competencies and standards of proficiency;
5) procedures used in selecting or developing test instruments;
6) public validation procedures;
7) grades tested and the testing schedule;
8) instructional strategies used to improve students' level of competency in relation to the standards;
9) how results of the program are to be used to meet the goals of the program, and dissemination procedures used in reporting results;
10) indicators of impact of existing programs;
11) financial costs of the program and funding sources;
12) provisions for special populations including minority students, bilingual students, and students with learning or physical handicaps.

Based on information collected during the site visits, a detailed description of each minimum competency testing program will be developed. To ensure their accuracy, these descriptions will be reviewed by the respective program administrators prior to publication. Additionally, NES will develop a Summary and Analysis Report describing current trends in minimum competency testing. The focus will not be the comparison or contrast of individual programs, but the synthesis of information about a variety of program types.

TASK 2: Develop Program Typology. The purpose of developing a typology for minimum competency programs is to provide a fixed framework within which programs can be characterized. The typology will permit categorization of programs on such dimensions as purposes, competency areas and grades assessed, time frames, and relationship to remedial programs. The typology will serve the additional function of ensuring that the design for the impact study to be developed will be sensitive to the variety of current program types in order that valid impact study information can be collected. While the nature of minimum competency testing programs will continue to develop, the typology should prove to be a useful resource for current program analysis.

TASK 3: Produce Program Resource Guides. The series of resource guides will identify alternative strategies and approaches to various aspects of minimum competency program development. Rather than advocate any single methodology, the guides will focus attention on the policy issues related to program design or modification. While the specific audiences for each guide will be surveyed for suggested topics, the following represent tentative topic areas:

1) defining competencies;
2) developing instructional strategies;
3) developing test instruments;
4) designing management strategies;
5) testing special populations (disadvantaged, minority, and handicapped students);
6) developing dissemination strategies;
7) conducting cost analysis studies;
8) measuring program impact;
9) involving special interest groups in planning programs.

For those administrators who must confront policy decisions about minimum competency testing, the guides will represent a comprehensive resource from which topic areas of particular concern may be selected. While each guide will function as an independent resource for a given area, the interrelationship of the guides will be presented in an initial guide that will provide an overview of all topic areas. A mechanism will be set up by NIE to disseminate the guides to state and local education agencies.

TASK 4: Design an Impact Study. The thrust of this task will be *planning* for the implementation of an impact study of minimum competency testing programs during Phase II of the project. In order to be sensitive to the variety of current programs, extensive input to the design of the study will be solicited from a variety of constituencies at both the state and local levels.

The development of the plan will include decisions about the scope of participation, the types of questions to be answered by the study, the specified student and school outcomes upon which impact should be assessed, the method of data collection, and a time frame for implementation. It is only through a process that emphasizes the input of all audiences that a feasible and useful impact study will be designed. Upon approval of the design, the impact study will be implemented in subsequent years as Phase II of the project.

Design of Phase 2

While the exact nature of the impact study is yet to be determined, plans are to involve a sample of state and local minimum competency testing programs. The study will, however, neither compare minimum competency programs in different states or local districts nor make relative judgments about the quality of individual programs. Rather, the goal will be to assess the representative impact of these programs in a way that will provide needed information about the results of minimum competency efforts. The study will be conducted over a three-year period in a manner that minimizes any resource burden on participating states and local districts and yet supplements data collection and analysis currently occurring in these states and districts.

Value of the Study

The outcomes of the study will be valuable to all educational decision-makers who must confront policy issues related to minimum competency testing. Its value will, therefore, extend to those undertaking initial efforts to develop programs and also to those reviewing the effects of already established programs. The study will be of benefit to anyone interested in an increased understanding of the variety of program types and strategies available, an awareness of the multi-faceted impact of these programs, and an opportunity to anticipate the minimum competency testing issues to be faced in the future.

Guidance for the Study

The success of the study will hinge upon participation by representatives of a wide variety of audiences connected with the minimum competency movement. These audiences will include legislators, state and local policy-makers and administrators, professional associations, special interest groups, teachers, and lay public advisory groups. The Council of Chief State School Officers, through its Committee on Evaluation and Information Systems, will be asked to facilitate contact with state and local agencies and will be informed of all project activities.

Input from a range of audiences will clarify the types of minimum competency activities currently in place and will identify those areas in which additional information and resources would be most helpful. Specifically, the identification of topic areas for program guides and the design of the impact study will reflect the input of those who are most likely to use the materials and results.

Throughout Phase I and Phase II of the study, NES and NIE will be assisted by an Advisory Committee composed of educators prominent in the field of minimum competency programs, one of whom is a representative from the Committee on Evaluation and Information Systems. The Advisory Committee will periodically react to the progress of the study, provide guidance in the development of the reports and resource guides, and make recommendations about the plan for the impact study.

SPONSOR:

National Institute of Education

This study is a project of NIE's Testing Team, Office of Testing, Assessment and Evaluation. The Testing Team sponsors research, development, and dissemination activities designed to improve the practice of testing in education. This study of minimum competency testing programs is part of a set of related activities investigating the role and impact of testing in education, and represents NIE's Contribution to the current debate on minimum competency testing.

CONTRACTOR:

National Evaluation Systems

National Evaluation Systems, Inc. (NES), located in Amherst, Massachusetts, is an organization which provides customized services in test development and test administration to state and local education departments, professional associations, federal agencies, and state licensing boards. Principal areas of NES expertise include: (1) the preparation of basic skills, functional literacy, and life skills tests to measure minimum competency for high school graduation and grade promotion; (2) the preparation of examinations and related materials in diverse occupational and professional specialty areas; (3) the administration of large-scale assessments for state, city, and regional education departments; and (4) the conduct of evaluations of educational programs and services.

The team approach taken at NES brings together on projects professionals experienced in such activities as conducting needs analyses, objective and item writing, test and instructional materials development, test administration, scoring, data processing and analysis, and reporting. Backed up by the more general NES capabilities of project management, training, and computer programming, these teams then provide a wide range of test-related materials development and administration services.

Federal I.D. No. 04-2524119

This project description will be revised and updated to include a more detailed description of the evaluation design on June 30, 1979.

PROJECT OFFICER AND PROJECT PERSONNEL

National Evaluating Systems, Inc.
30 Gatehouse Road
Amherst, Massachusetts 01002

PROJECT PERSONNEL:

William Gorth, Ph.D.
Project Director

Sherry Ann Rubinstein, Ph.D
Project Coordinator

Marcy R. Perkins, Ph.D.
Senior Research Associate

Mary Tobin
Research Associate

STAFF:

Richard Allan, Ed.D.
Paula M. Nassif, Ed.D.
Paul Pinsky, Ph.D.
Peter Schriber, Ed.D.
Claudia Sheehan

National Institute of Education
1200 Nineteenth Street, N.W.
Washington, D. C. 20208

PROJECT OFFICER:

Judith S. Shoemaker, Ph.D.
Senior Associate and
Team Leader for Testing
Office of Testing, Assessment
and Evaluation
(202) 254-6271

LIAISON WITH COUNCIL OF
CHIEF STATE SCHOOL
OFFICERS:

Gerald Bracey, Ph.D., Director
Division of Research, Evaluation
and Testing
Virginia Department of Education

U.S. Department of Health, Education and Welfare
Joseph A. Califano, Jr. Secretary
Mary F. Berry, Assistant Secretary for Education

National Institute of Education
Patricia Albjerg Graham, Director
Sylvia Scribner, Associate Director
for Teaching and Learning

1200 Nineteenth Street, N.W.
Washington, D. C. 20208

ADVISORY BOARD

Gerald Cavanaugh, Principal
Kaiser Elementary School
Denver, Colorado

James Findley, Vice-Principal
Westside High School
Omaha, Nebraska

Thomas H. Fisher, Ph.D., Director
Student Assessment Section
Florida Department of Education

Donald Henderson, Admin. Assistant
Educational Svcs., Secondary Education
Gary Community School Corp., IN

George Madaus, Ph.D.
Professor of Education
Boston College, MA.

Jason Millman, Ph.D.
Professor of Education
Cornell University, N.Y.

Gordon Ascher, Ph.D.
Associate Superintendent
Educational Program Audit Division
Oregon Department of Education

ATTACHMENT 4

Handout concerning Draft Standards for Educational Evaluation.

This attachment was removed prior to submission to the ERIC Document Reproduction Service.

For further information contact:

Daniel L. Stufflebeam
Chair, Joint Committee on Standards
for Educational Evaluation
Evaluation Center
Western Michigan University
Kalamazoo, MI 49008

Why Should Some Tests Be Secure?[1]

John Fremer and Alice J. Irby
Educational Testing Service

THE QUESTION -- WHY SHOULD SOME TESTS BE SECURE?

Before providing an answer to the question -- Why should some tests be secure? -- it will be useful to explain what is meant by test security and to indicate the nature of the relationship between test security and the purpose for testing. Secure testing programs are those in which the individual test forms are administered under professionally supervised and standard conditions. The initial development and subsequent use of the test questions in secure tests is carried out so that no test taker enjoys an advantage over others through prior familiarity with the specific questions included in the test. The following are examples of such tests used in the field of education: the Scholastic Aptitude Test of the College Board, the American College Testing Program, the Medical College Admissions Tests, and the Tests of General Educational Development (High School Equivalency) of the American Council on Education.

Test security is an essential component of these state and national testing programs because the performance of test takers will contribute to selection decisions. Under these circumstances, principles of fairness and equity require equivalent conditions for all test takers. Only secure testing programs permit such fair and equitable treatment of test

---

[1]Preliminary Paper presented at the Spring Conference of the National Consortium on Testing, Arlington, Virginia, April 1979.

takers. This conclusion has been reached independently by developers of critical testing programs not only in the United States, but in other countries.

Although this paper argues the value of test security for certain testing programs that are associated with selection decisions about individuals, there are many testing contexts in which security is not a critical issue. The largest use made of educational tests is in the classroom setting. In this context many teachers make use of "take-home" examinations that permit student collaboration on test questions. Many other teachers return examination copies routinely for review and guidance. In general school achievement batteries of test publishers are readily available for teacher review and analysis. Example: the State of New Jersey publishes its educational assessment instruments each year after the examinations are administered.

It is clear that when examinations are closely linked to an ongoing instructional process or, as in the New Jersey program where it is judged important to call attention to results on specific questions, individual test forms are published. For these examinations test security after administration and, in some cases, even during the administration is not a primary concern.

OBJECTIVES OF TEST SECURITY

1. Fairness -- Test questions would be invalidated for use in future secure tests on the grounds of fairness whenever some test takers will

have had a prior opportunity to study these questions and others will not have had such an opportunity. Theoretically, there are two ways to deal with the problem of fairness of access to test questions: Giving all test takers the opportunity to see all questions, or keeping all questions secure so that each test taker encounters the questions for the first time on the day of the examination.

Giving all test takers equal access to questions is not a realistic alternative. There is, first of all, the wide variety of circumstances in which test takers function -- some are full-time students, others are in full-time careers. Some test takers live in situations that would permit intensive study and review of test questions, others would have little or no time to devote to such an activity. Yet when test results play a part in admission or employment decisions about individuals, it is essential that testing be conducted in a manner that is impartial. Keeping test forms secure insures that each test taker is taking the same or an equivalent test under the same conditions.

2. Meaningfulness of Test Results -- Only when all test takers encounter testing materials for the first time under similar examination conditions is it reasonable to use performance on the sample of test questions as a basis for making inferences to performance on the larger body of material from which the test is drawn. When some test takers

have had prior exposure to specific test questions and others have not, there is no way of adjusting one's inferences from the test score to take into account differential exposure to the materials. It is for this reason that test security is critical to fair and objective assessment.

Colleges, universities, professional bodies, and other institutions that use test scores as one element in making decisions about individuals need a fair and objective basis for making inferences about the knowledge and skills of the test taker. Responses to the sample of questions included on a test are important as a basis for generalization to the larger domain that is represented. It is rare that the specific set of questions administered constitutes the full domain of interest to the user of test scores.

3. Continuity of Test Score Scales -- By standardizing testing conditions and carrying out careful test development procedures and statistical equating studies, it is possible to place scores from different editions of a particular test on the same score scale. As test users acquire familiarity with the resulting scores and appropriate normative and interpretive material, they can make appropriate inferences from test scores without reference to the particular test form that was used.

The equating procedures currently in use in secure testing programs of Educational Testing Service (ETS) require the comparison of student

performance on a new set of questions with their performance on questions previously administered within the testing program. One widely used method of carrying out this comparison is to administer, as part of a new test, a sample of questions that had been previously administered as part of the same program. If all such previously administered questions had already been released, it would no longer be possible to treat these questions as identical for purposes of establishing the relationship between the new questions and the previously administered questions. Releasing items would limit flexibility in maintaining studies of the stability of a test scale.

The present equating methods grew out of years of research. Developing alternative methods is extremely complex. In addition to requiring research and statistical work, it would probably be necessary to obtain the statistical data needed from a test-taking population different from the operational test-taking group. Not only would there be considerable expense associated with such special testing, it would be necessary to take into account any differences in the population used for these analyses and the group that participated in the regular testing program. There would be substantial risks that the results would be subject to greater error than existing methods.

4. Pretesting of New Test Questions -- Potential new test questions for aptitude tests are frequently administered to test takers along with operational test questions. Evaluation data are thus obtained for highly

motivated and representative samples of students. These data then serve as the basis for preparing high quality items that have well-defined content and statistical properties and that have been pretested on similar populations. The loss of such opportunities for pretesting could affect substantially the quality of new test forms.

5. Reuse of Test Questions -- The maintenance of test security permits ETS and test program sponsors to reuse individual test questions and test forms. This makes it possible to offer a variety of different regular test administration dates as well as special and "make-up" test administration dates for students who, for religious or other reasons, cannot sit for examinations at the regular dates.[1] It also keeps costs lower than they would be otherwise.

Considerable professional time is invested in developing, reviewing, and editing items, and challenging ambiguous test questions. The involvement of educators and other professionals from a wide range of disciplines and specialties is needed to achieve accuracy, clarity and appropriateness. It would be wasteful and uneconomical to limit the use of test questions to a single occasion.

6. Discourage Teaching to the Test -- One possible consequence of the release of secure test items would be an increase in the use of drill

---

[1]For example, the Scholastic Aptitude Test of the College Board is administered nationally on six weekends in the course of a year and in five states on a seventh weekend. On each weekend there is both a Saturday and a Sunday administration. To cover situations such as severe weather conditions and other emergencies, special local administrations are arranged at other times during the year. Additional test forms are available for the testing of handicapped students.

and practice on test questions as a substitute for productive educational activities. Although the evidence that has been accumulated suggests that gains in test score from drilling on test questions are likely to be insignificant, test takers and others may view the release of many items and tests as encouraging such drill. ETS would not want to encourage teaching to a test in lieu of important instructional activities.

WHAT ONE NEEDS TO KNOW ABOUT SECURE TESTS

Assuming the importance of test security for certain kinds of examinations, the critical question then becomes, "What information is needed by and important to a) the candidate and b) the users of the scores?"

From the perspective of the test taker the following needs for information are of major importance:

test content -- What is the test like? What does it measure?

sample questions -- What form will the questions take? Do I understand what I am supposed to do?

sample tests -- If I take the test at my usual pace of work will I have enough time? Is it fair or tricky? How does the test match my education and experience?

test use -- How do institutions use the test scores? What are the consequences for me of earning low, average, or high scores?

test interpretation -- What meaning should I attach to the scores? What interpretive data are available to help make sense out of them?

limitations -- What are the limitations of the test in drawing inferences? What uses would be appropriate or inappropriate?

From the perspective of the user (the institution or agency) the same information as above as well as the following are important:

adequacy of development process -- What evidence is there that the examination was developed by qualified individuals using sound technical procedures?

validity information -- What evidence exists regarding the appropriateness of the test for the purpose for which I want to use it?

reference data -- What normative or other reference data are available that relates to my setting? The local population? A national population?

quality of scoring and reporting process -- How effective are the quality control procedures that protect the accuracy of the scores that I receive?

The testing programs of ETS and other testing organizations need to fulfill these information needs. Test program sponsors, institutional users, professional associations, test takers, and others have strong grounds for complaint and criticism when adequate evidence and data are not available. We feel that ETS programs can stand very close scrutiny against these criteria. Of course it is the right of each individual or institution to form a judgment on any one of these criteria or on the set of results for a particular testing program.

NEED TO PROVIDE APPROPRIATE INFORMATION ABOUT SECURE TESTS

The importance of decisions based on secure tests makes it essential that test candidates and test score users have adequate information on which to base judgments and actions about tests and test scores. Providing access to sample test forms is one useful component of such information. Clearly, additional methods of providing needed information are called for.

The concerns that have been raised about access to, and available information about, secure tests deserve serious attention by ETS and other testing organizations. We need to determine what use is now being made of the information that is already available. It may be that we will determine that there are information needs that are not being met and which will require expanded development or dissemination of sample test materials, descriptions, and statistical and research data.

WHAT ABOUT THE RELEASE OF ALL TEST ITEMS?

Among the millions of candidates who take secure examinations in the course of a year, there is little evidence of significant demand for or need for the release of the items themselves. Moreover, there is no clear idea of what the gain would be, expecially when juxtaposed against the need to maintain impartiality and fairness by keeping items secure. For instance, how much more would a student know about himself or his performance in relation to peers from examining items on a specific form of an aptitude test than he would from a) understanding the total score or subscores in relation to their use and b) examining

sample items and tests? Many of the tests in national testing programs involve aptitude tests which measure skills that develop over a large number of years in a variety of formal and informal educational settings. Merely studying the actual aptitude test questions that were administerd would be of little to no educational value.

While there may be some marginal gain for some candidates, the offsetting loss in terms of the importance of security to large numbers of other candidates may be much greater. In view of this doubt as to the value of releasing all items, the benefits do not seem to outweigh the drawbacks which are:

1. Problems of Interpretation -- It would be almost impossible for the average recipient of question-by-question test results to interpret them adequately without background information regarding the test specifications, the difficulty level of the questions, and the relationship of individual test questions to the test as a whole. It would clearly be important to offer extensive complex background material on the test questions and the test as a whole in an effort to avoid misinterpretation. Such information tends to be technical in nature and not readily converted into terms familiar to the candidates. It is doubtful that the additional information made available would contribute significantly to the learning process of the individual. Indeed, an excessive focus on individual items could be misleading to the student.

2. Depletion of Pools of High Quality Items -- The test making process at Educational Testing Service is an extraordinarily open one involving thousands of educators and other professionals and specialists drawn from across the nation. All tests prepared by ETS are subject to extensive external professional review. Virtually all tests are developed in collaboration with external committees of examiners with a membership that gradually changes over time to permit broad representation of levels and types of institutions, subject matter, academic, or occupational fields, and various ethnic and minority groups.

In some fields, it may be extremely difficult to generate the quantity of items necessary for pretesting and the maintenance of quality now exhibited in the tests. ETS, through hiring staff or consultants to meet this need, would endeavor to maintain quality. However, some tests, especially achievement tests might have to be offered on fewer occasions, resulting in fewer administrations and reduced access for candidates. For example, the examination for real estate licensing is administered twice a month. Generating enough items for new forms, given the limited domains covered, would be so difficult that the number of administrations would probably be reduced.

3. Cost -- The costs of developing and administering examinations that were published after administration would increase substantially over the costs associated with secure testing. The amount of the increase

would depend on the number of test administrations, the pretesting of new items, the costs of equating, the extent of item review, the availability of pretest populations for field testing, and the generation of new items.

In addition to the costs associated with the writing, review, and evaluation of additional test questions for new test forms, using items only once would add to test development costs because it would no longer be appropriate to equate new tests to previously used tests through the readministration of questions used on the earlier test forms. Yet an equating process is essential to insure comparability of test scores from test form to test form.

The task of pretesting or trying out new test questions also entails cost and test quality implications. When potential new test questions can be tried out along with operational test questions, highly motivated and representative samples of students provide reliable cost effective data at minimum cost, as a basis for preparing future test forms.

A wide variety of expenses are associated with the development of questions for secure tests, including costs related to the planning and creation of test specifications or blueprints, initial writing of questions, review and editing, statistical analyses, and the final selection, arrangement, and further review of complete test forms. Since the number of questions required to achieve reliable and valid measurement varies as a function of the purpose of testing, the nature of the field tested,

and the nature of the test-taking population, no single figure can adequately characterize the expenses involved. At the present time, the cost of developing questions for a single secure test administered by ETS falls within a range of $45,000 to $125,000, exclusive of supporting research and development activities.[1]

THE PUBLIC'S NEED TO KNOW

Although the maintenance of test security is based on the need for fairness and equity in testing, the concept of security is apparently easy for individuals to link, inappropriately we believe, with ideas of secrecy and mystery. Critics of testing in general, and of ETS as an institution, paint pictures of a small group of insensitive and basely motivated test makers carrying out secret and almost certainly evil deeds under an impenetrable shroud.

In view of this, a word about a test publisher's obligation to the public may be useful. Educational Testing Service, like many organizations, has multiple constituencies: its clients and sponsors who are the primary consumers of secure tests, the candidates, educators such as guidance counselors, the educational and psychological research community, and professional measurement associations. The crux of the matter is the balancing of these interests. How can test publishers meet what may appear to be conflicting needs of institutions and associations on the one hand, and the public's need to know?

---

[1] Keep in mind that in a program such as the Admissions Testing Program, the Scholastic Aptitude Test with its two components is administered a number of times during a year and fifteen achievement tests are also administered on multiple dates.

Educational Testing Service meets this general need to know in several ways. First, it does so by publishing sample tests, descriptive information, interpretive brochures, technical manuals, and research reports. Secondly, it invites external review of its secure tests, both editorial review by subject-matter specialists and members of ethnic groups and technical review by reviewers for the Mental Measurements Yearbook and other publications. Persons and agencies planning to use specific examinations are invited to review the tests for usefulness and adequacy. In some cases, courts have ordered appropriately, the review of tests by qualified professional reviewers.

External review, however thorough and sound is no substitute for internal surveillance. Independently of its clients, who have standards of their own, ETS has produced its own Principles, Policies and Procedural Guidelines Regarding ETS Products and Services against which its programs are monitored and audited. Thus, ETS has made available to its clients, candidates, and the public the standards against which it conducts its activities.

Finally, authority concerning its operation is dispersed among sponsors, committees, professional associations, staff, and institutional users. While ETS' approach to building secure tests may be more time consuming than other possible methods, the involvement of many people from many institutional environments is compatible with the diversity and decentralization of the educational society ETS serves.

SUMMARY

Without doubt, the publication of test questions and answers is appropriate for non-secure tests used for diagnosis or placement of students and in those instances in which professional counseling is available. The classroom teacher, for example, can use the question-by-question results for his or her class to help plan appropriate instruction.

However, the calls for publication of secure test items fail to recognize the values and purposes of this approach to testing. We believe that the call for test publication under such seductive and misleading labels as "open testing" and "truth in testing" merely detracts from consideration of important issues related to the development, interpretation, and use of tests. The issue of release of test items may be valuable if it contributes to careful examination of the purposes underlying the arrangements for testing that are now in place and to an increasing awareness and understanding of tests by users, candidates and the public generally.